SURVEY: A PLATFORM TO EXPLORE EMERGING VALUE PERSPECTIVES IN ALMS' BEHAVIORS

A PREPRINT

Steven Bickely, Ho Fai Chan, Bang Dao, Benno Torgler, Son Tran

Panalogy Lab
Brisbane, 4000
steven.bickley@panalogy-lab.com

July 3, 2023

ABSTRACT

This paper presents our work on Survey, a platform for analyzing augmented language models's (ALMs) emergent alignment behaviours through their value perspectives.

***K*eywords** template • demo

1 Introduction

Our new platform, Survey, aims to explore emerging value perspectives in Augmented Language Models (ALMs) [1] behaviors through decision premises and contextuality in complex social settings. **[why complex social settings?]**. Our approach to exploring ALM's behaviors, and their value perspectives, is based on presenting the models with complex sets of survey questions then analyzing their responses under in various contexts. **[why do we use a survey-based approach to study value alignment? is this a proven methods? ref to seminar papers?]**

The first key aspect we address is the question of why understanding ALM's value perspectives matters in alignment research. By examining the context-dependent values that drive ALMs' behaviours, we gain valuable insights into the ethical and moral frameworks and traces that shape their decision-making processes (as well as our own). These insights can help us align ALMs with (context-specific) human values and promote responsible AI development [2].

Defining measures of values is another crucial element of our platform, Survey. We recognize that values are multi-faceted and subjective, making their quantification challenging. Through rigorous research and analysis, we aim to develop and translate robust methodologies to define and measure values in the context of ALMs. This will provide a solid foundation for studying and understanding their behavior patterns in complex social settings.

To effectively explore emerging value perspectives in ALMs' behaviors, we employ state-of-the-art (SOTA) ALM simulation techniques. By simulating ALMs in realistic scenarios, we can observe their decision-making processes and identify any evolving biases or value systems. This knowledge is crucial for staying ahead of potential risks and ensuring the responsible development of AI technologies.

The potential applications of our platform extend beyond the fields of behavioural economics, cognitive psychology, or market research. We envision leveraging the insights gained to help drive social AI alignment and design in research, industry and community alike. By aligning AI systems with human values, we can (co-)create social AI systems that complement and enhance lives, benefiting society as a whole.

In conclusion, our platform Survey offers a unique opportunity to explore and understand the emerging value perspectives in ALMs' behaviors. By addressing the questions of why world values matter, defining measures of values, and simulating ALMs in complex social settings, we can unlock insights that will help shape the future of AI development. With a focus on social AI alignment and design, we aim to create long-lived, complex social AI agents and systems that align with human values, promoting a responsible and beneficial integration of AI into industry and society.

2 Background

With the emergent capabilities of Large Language Models (LLMs) and the ongoing development of Augmented Language Models (ALMs), understanding the underlying premises that govern their behaviors and decision-making processes becomes increasingly crucial, especially when they operate in complex social settings and have real-world impact. This understanding is essential to ensure the alignment of social AI with human values.

ALMs, which are essentially like advanced versions of pre-trained LLMs like gpt-4, incorporate various elements such as retrieval plug-ins, different learning techniques (few-shot), diverse prompting methods (such as chain-of-thought, self-model, and contextual prompts), functional coding, and integration with other modalities like voice, vision, and sound [1]. Additionally, future iterations of ALMs are expected to incorporate different AI techniques such as reinforcement learning and symbolic logic, enhancing their knowledge organisation, reasoning, and learning capabilities.

Researchers have been quick to recognize the potential of LLMs as valuable tools to study and probe the human mind and society[6], given their training on vast amounts of human data and their ability to generate human-like text. Others have discussed their potential in simulating human subjects [9]. Consequently, researchers from various disciplines such as behavioral economics, cognitive psychology, social psychology, linguistics, and others have proposed different tools, surveys, and methodologies to investigate LLMs' behavior and decision-making processes. However, the procedures and tuning of LLMs (e.g., temperature, context window, prompt context and structure) for judgement and evaluation of alignment are not yet standardized or consistently applied across studies. Moreover, digital literacy and programming skills continue to present significant obstacles for many researchers, particularly those in behavioral economics and the social sciences.

Considering the fast-paced nature of research and development in AI at the moment, it is essential to also extend our focus beyond pre-trained LLMs and consider the emergent capabilities and value systems of ALMs within various different social contexts. ALMs are increasingly augmented with additional tools and various prompting techniques, spanning different context windows and incorporating other modalities. Furthermore, these ALMs are now actively performing real-world actions. Calling a tool in the context of ALMs often involves having an impact on the virtual or physical world and observing the resulting effects, which are typically integrated into the ALM's ongoing context. Moreover, ALMs are increasingly engaging in delegate actions such as carrying out transactions on our behalf or responding to customer queries and emails in human-like ways.

By acknowledging the advancements in ALMs and the complex nature of their interactions with the world, we can gain a comprehensive understanding of the premises underlying their behaviors and decision-making processes. This knowledge is crucial for ensuring the development of responsible and aligned social AI systems that reflect human values for the benefit of all humankind. To facilitate this exploration in an easy and intuitive manner, we have developed a social science research platform called Survey . It empowers researchers to investigate the behaviors and decision-making of LLMs and ALMs in a robust and systematic way, using an easy-to-use, click-and-play user interface

3 Framework

To be continued...

4 Some Empirical Findings

[Here we can add a section showing results of using Survey to answer the world-values questions. Thus world-values survey is just only one of many use cases for Survey.]

5 Limitations and Future Avenues

Whilst we do begin with Survey to explore the emergent values in ALMs behaviours and decision making in a robust and systematic way, the augmentation of LLMs in this current proof-of-concept version of the platform. We have not yet explored other computational and decision-making techniques like reinforcement learning, or the utility of other modalities. This is a key future avenue for this platform.

A key limitation of the social context of our simulation is that context is provided in a randomized manner, and using a generic story template to build up the context of the agent (e.g., you are <AGE>, a specialist in <OCCUPATION>, and operate in <LOCATION>). We do not draw on or attempt to simulate human subjects from demographic backgrounds of past survey respondents, as in e.g., [10]

As an initial proof-of-concept, we have naturally had to limit the number of context variables available in the Survey platform. Here are some dot points indicating the social context variables that are not currently implemented in our platform, but are planned future additions:

Employment status (employed full time, employed part time, employed casual, self-employed, unemployed, studying part time and working full time, studying full time and working part time, studying full time and not working)

- Rural vs urban residence
- Mental health status
- Cognitive style (e.g., analytical, intuitive, practical, creative)
- Marital/family status
- Occupational industry
- Immigration status
- Veteran status
- Home ownership vs. renting
- Food security status

References

[1] G. Mialon *et al.*, "Augmented Language Models: a Survey."

[2] A. Tamkin, M. Brundage, J. Clark, and D. Ganguli, "Understanding the capabilities, limitations, and societal impact of large language models," doi: 10.48550/arXiv.2102.02503.

[3] A. Arora, L.-A. Kaffee, and I. Augenstein, "Probing Pre-Trained Language Models for Cross-Cultural Differences in Values."

[4] M. Binz and E. Schulz, "Using cognitive psychology to understand GPT-3," *Proceedings of the National Academy of Sciences*, vol. 120, no. 6, p. e2218523120, Feb. 2023, doi: 10.1073/pnas.2218523120.

[5] A. Korinek, "Language Models and Cognitive Automation for Economic Research," 2023.

[6] M. Miotto, N. Rossberg, and B. Kleinberg, "Who is GPT-3? An Exploration of Personality, Values and Demographics."

[7] G. Aher, R. I. Arriaga, and A. T. Kalai, "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies."

[8] J. J. Horton, "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?"

[9] J. S. Park, L. Popowski, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Social Simulacra: Creating Populated Prototypes for Social Computing Systems."

[10] L. P. Argyle *et al.*, "AI Chat Assistants can Improve Conversations about Divisive Topics."